Variant Graph Convolutional Networks for Skeleton-Based Hand Action Recognition

Khin Sabai Htwe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
khinsabaihtwe@toki.waseda.jp

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Graph convolutional network is widely used in skeleton-based applications such as action recognition. In this paper, a Variant Graph Convolutional Network (VGCN) is proposed to learn not to be constrained of the physical connections of hand structure since a predefined fixed graph structure lacks of flexibility to capture variance and different actions. With experiment on our hand actions skeleton dataset, the proposed method outperform with significance accuracy to the conventional ones.**

Keywords—variant joints, skeleton information, hand action, recognition, graph convolutional network

I. INTRODUCTION

Skeleton-based applications have been widely investigated and used due to their robust adaptability to the varying lighting conditions and complicated background. Skeleton sequences of hand joint coordinates can be easily obtained from the robust pose estimation methods such as OpenPose [1] which can provide x, y coordinates and confidence of each joint.

Recent works of action recognition had most widely used skeleton data representing as vector sequences and graphs in deep-learning methods to classify the actions. In proposed paper [2], the authors present multilayer LSTM networks on geometric features for action recognition. They adopt a typical human skeleton model with 16 joints, any two of joints form a line, and any three of joints form a plane for geometric features. Since the number of pairwise combination of joints, lines and planes is extremely large, it could be time consuming. Therefore, they select important lines and planes to make edge connection of the movements of action. In Spatial-Temporal Graph Convolutional Networks (ST-GCN) [3], they proposed a partition strategy based on the location of the joints and the characteristics of the movement of the human body. Since they define edges according to the connectivity of human body structure, it may neglect joints which can become rich information to identify actions. Based on observed conventional works, we analyzed to know that edge selection plays an important role and found that it may influence on the recognition accuracy.

With this in mind, we propose a variant graph convolutional network (VGCN) to learn not to be constrained of the physical connections of hand structure since a predefined graph with fixed structure lacks the flexibility to capture the variance and different actions.

II. PROPOSED METHOD

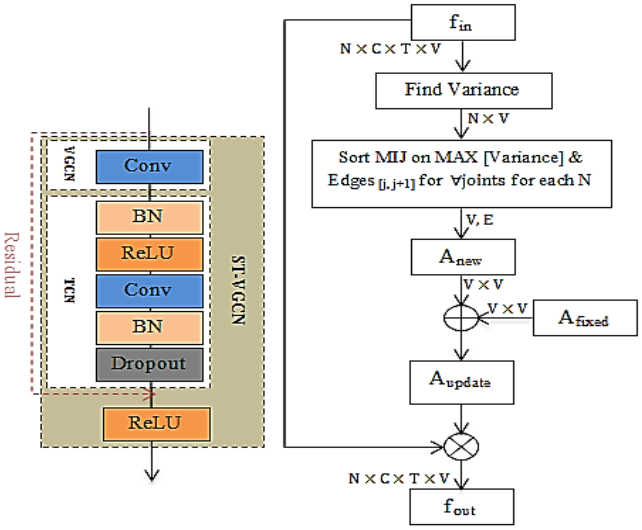
In case of this issue, we present an approach with selecting most variant joints to define features for classifying actions. We use ST-GCN [3] as baseline method. In contrast to their method, our VGCN network is considered by joint constraints on different actions as mentioned in the following:

A. Joints Influence for Edge Connections

First, statistical analysis is used to evaluate which joints has more varying or utilization in each action. It can infer that the more the variance of joints, the more the important of that joints for each action. Those are called Most Influence Joints (MIJ). Based on these variant joints, new edge connections are defined including weak connections on each action.

B. Updating Graph based on Edge Connections

According to new edge connections, an adjacency matrix is learned by combining to old one not to be constrained of the physical connections of hand structure. And then learned features are extracted by batch multiplying of that adjacency matrix and input features. The flow of graph is shown in the following figure Fig.1.



(a) (b)

Fig. 1. Flow of Variant Graph Convolutional Network, (a) ST-VGCN, (b) Details of VGCN

After learning new edge connections in graph convolution, the extracted features of ST-VGCN networks can be predicted by softmax function with weak connections for each action.

III. EXPERIMENT

A. Hand Action Skeleton Dataset

Since the original hand action dataset [4] which captured of 18 different subjects has no skeleton data, hand skeleton data of each action is extracted using our previous work [5] which can provide x, y coordinate and confidence of each joint. These x, y and confidence are normalized and used as 3 input channels to the ST-VGCN network as shown in Fig. 2. In this skeleton dataset, the 120 videos and 24 videos are used for training and testing respectively for 8 actions such

as Blue, Clicking, Double-clicking, Green, J-Sign, Scissors, Swiping and Yellow. For each video, there are 240 frames at a frame rate of 30 fps with resolution 640x480 pixels.

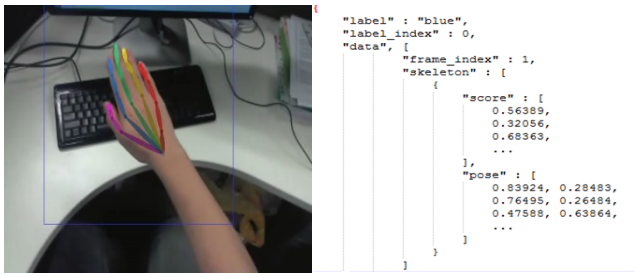


Fig. 2. Example of the "Blue" action and its skeleton data

B. Initialization Settings

To start with, joints and edge links should be initialized according to hand pose structure likes Fig. 3. Therefore, a baseline edge includes with self and adjacent links to construct an adjacency matrix. Thereafter, a learned adjacency matrix is obtained by updating new edge connections on different actions.

Fig. 3. Hand Skeleton Joints [1]

C. Variant Graphs

Here in Fig. 4. are some results of learned adjacency matrix of third input.

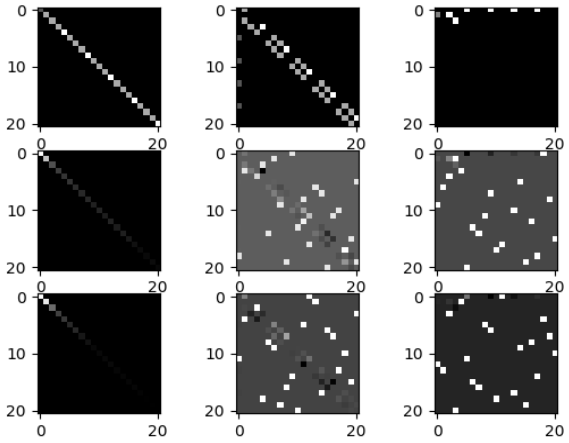


Fig. 4. The first row is the original fixed matrix and the rest two are learned adjacency matrix of 1st and 2nd layer of third input

To make it more clearly, the first row is for a fixed graph and the rest rows are variant graphs for 1st and 2nd layer, respectively. As for columns, the first one is symmetrical links and the rest two are inward and outward links, respectively for each edge connection. As shown in Fig. 4, learned adjacency matrix can carry out more information beyond hand pose structure.

D. Training Details

In training phase for ST-VGCN model, three skeleton coordinate data is fed to the network as three input channels. There are total 10 ST-VGCN (as shown in Fig. 1(b)) layers including with spatial and temporal unit by utilizing Adam optimizer, learning rate 0.01, batch size 8, and 100 epochs for fair comparison with baseline ST-GCN method.

IV. Results and Discussion

To validate our method, different test samples are used in model. As the learned adjacency matrix can carry out the rich information including weak connections by evaluating variance of actions, the extracted features are accumulated by multiplying that matrix and input samples and have more clarifications to classify actions using Softmax layer. After that these probabilities are used to calculate which class label has the highest accuracy and lowest loss. To represent probabilities to class labels, the cross-entropy loss function is used as in equation (1):

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i), for\ n\ classes \quad (1)$$

Where t_i is the truth label and p_i is the softmax probability for the i^{th} class. The smaller the loss, the better the model is. Accuracy measures the proportion of the model predicted label matches the target label. Contrast to the baseline ST-GCN, our approach improve the performance of hand action recognition rates on our dataset in TABLE I.

TABLE I. Comparisons of the Validation Accuracy

Methods	Accuracy (%)	Loss
ST-GCN [3]	70.83	0.726
ST-VGCN (Ours)	**79.17**	**0.716**

V. Conclusions

In this paper, we summarize the previous works of action recognitions and analyze the relations of edges and its important role. Based on fluctuation of analysis results, we propose the VGCN network to choose edge connections by finding variance of each joint to decide which of them is the most varying for action and by learning adjacency matrix to update appropriate graph beyond the hand physical structure. This idea makes a complement to the state of the art method and further enhances the performance. For further works, geometric data will be considered to be more robust features to classify the actions.

References

[1] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1145-1153, July 2017.

[2] S.Zhang, X.Liu and J.Xiao, "On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks", the IEEE Winter Conference on Applications of Computer Vision (WACV), pp.148-157, March 2017.

[3] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", Proceedings of the Thirty-Second {AAAI} Conference on Artifical Intelligence (AAAI-18), pp. 7444-7452, February 2018.

[4] C. Xu, L. N. Govindarajan and L. Cheng, "Hand Action Detection from Ego-centric Depth Sequences with Error-Correcting Hough Transform", Pattern Recognition, vol 72, pp. 494-503, August 2017.

[5] K. S. Htwe, T. Ishikawa, and H.Watanabe, "Improving Detection of Hand Joints in RGB images Using Maximum Confidence Values", 2018 ITE Winter Annual Convention, 12D-5, December 2018.